White Paper Report

Report ID: 109287

Application Number: HD-51636-13

Project Director: Susan Weiss (sweiss@jhu.edu)

Institution: Johns Hopkins University

Reporting Period: 5/1/2013-10/31/2015

Report Due: 1/31/2016

Date Submitted: 2/1/2016

NEH WHITE PAPER

White Paper

Report ID: 113804

Application Number: HD-51636-13

Project Director: Susan Weiss (sweiss@jhu.edu) Institution: Johns Hopkins University

Report ID: 109289

Application Number: HD-51636-13

Project Director: Susan Weiss (sweiss@jhu.edu) Institution: Johns Hopkins University

Reporting Period: 11/1/2014-1/31/2016

Report Due: 1/31/2016

Date Submitted: 1//31/2016

Digital Prosopography for Renaissance Musicians: Discovery of Social and Professional Networks

Susan Weiss and Ichiro Fujinaga

Johns Hopkins University 31 January 2016

a. Project Activities

- **Software Development**

In October 2015, following two semesters of work, our Johns Hopkins graduate students completed creating a web-based work-flow environment with an easy-to-use flexible editor of RDF data (quads) and a simple interface to SPARQL At that point, because of a number of challenges (code spread over a number of different repositories, a number of coders, some of them not fluent in English who were documenting in Chinese), and, as a result of the bugs and flaws created by these complexities, the principals sought the professional assistance of a former graduate student at the University of Maryland, Max Morawski.

What we discovered was that the dependencies had to be figured out in order to rebuild the project from scratch. Max took on the following:

1. A web interface for uploading text files to be processed.
2. NLP programs in Java (OpenIE, nlptools, etc.) that extract relationships (subject-predicate-object) from the text.
3. Import the resulting triples into the RDF Editor.

He then moved all code in one repository and began restructuring to bring it more in line with standard practices. He also made it possible to work on the entire project with little effort using an Eclipse project and Maven. In the third and fourth week of his one month on the project, he was able to make some critical repairs such as fixing security flaws, and made some cosmetic and developer changes such as the removal of extraneous files (over a dozen), and the correction of typos. He also added new features including passing the RDF triples to the RDF editor after inspection and set Open IE libraries to preload. As a result of his work, upload.html now does the actual RDF parsing and file uploading is

simpler and the text of the files is displayed. The website is hosted on Amazon Web Service including a page for uploading and deleting files, a code for parsing text into RDF triples using Open IE (not attached to the website) and an RDF event editor

- **Project Activities: Knowledge Mobilization, Publicizing**

All throughout the process our findings have been available on (http://www.humanhistoryproject.ca). In addition, Fujinaga and Weiss gave a poster presentation entitled: "Digital Prosopography of Renaissance Musicians: Discovery of social and professional network" at the Annual Meeting of the American Musicological Society / Society for Music Theory in Milwaukee on 6 November 2014. We had excellent feedback. We have broadened our base of musicologists in an effort to begin to answer queries. We were also awarded a no-cost extension on the project through to next October 2015 in an effort to finish programming, mine data and continue software development. Our team is looking at ways of solving the problem of teaching the machine to recognize predicates. We are currently experimenting with NLP using reVerb. Our presentation was well-received at the RSA meeting in Berlin in March 2015 where we had queries from musicologists and from scholars in other disciplines. We presented our latest results at the International Association of Music Libraries, Archives and Documentation Centres (IAML) / International Musicological Society (IMS) Congress in New York City on 22 June 2015 at the Juilliard School. Several questions were raised including one about foreign languages. Coincidentally, at that same meeting, we met researchers from abroad who expressed interest in partnering with us in future symposia. Another of our abstracts was submitted and accepted for the Conference on Interdisciplinary Musicology "Imagination in Music, Shanghai, China, 27–29 November.

(see Appendix). In December, Professor Fujinaga hosted a workshop— The second international workshop on Human History Project: Natural language processing and big data— at the Schulich School of Music at McGill University in Montréal (http://www.cirmmt.org/activities/workshops/research/human_history_project_2/hpp). This workshop was organized by Research Axis 2 of CIRMMT (Centre for Interdisciplinary Research on Music Media and Technology) and was the second workshop on a large project called Human History Project, which aims to build a distributed international database of documented human history using Natural Language Processing tools and Linked Open Data to model historical data. The speakers were:

Jason Boyd, Department of English, Ryerson University
Serge ter Braake, Faculties of Humanities, University of Amsterdam
Ichiro Fujinaga, CIRMMT, Schulich School of Music, McGill University
Sergio Oramas, Music Technology Group, Universitat Pompeu Fabra
Matthew Milner, Department of History, McGill University
Susan Forscher Weiss, Peabody Conservatory of Music, Johns Hopkins University.
Following the workshop the speakers met to map out next steps.

In addition to the conference presentations, we gave seven invited talks regarding this project:

University of Kyoto, Kyoto, Japan. 28 January 2016. *Research Program of the Distributed Digital Music Archives and Libraries Laboratory.*

Japan Advanced Institute of Science and Technology (JAIST), Kanazawa, Japan. 21 January 2016. *Human History Project.*

School of Information Studies Research Seminar, McGill University, Montreal. 13 November 2015. *The Research Program of the Distributed Digital Music Archives and Libraries Laboratory.*

CCS Colloquium, University of Virginia. 2 October 2015. *10 Years of Innovation in Music Digitization & Dissemination.*

Oxford eResearch Centre Seminar. 28 July 2015. *Research Program of the Distributed Digital Music Archives and Libraries Laboratory.*

McGill Digital Humanities Work in Progress Talk, McGill University, Montreal. 9 February 2014. *Human History Project: Storing documented human history into a database.*

École Polytechnique de Montréal, Montreal. 19 November 2013. *Digital Prosopography of Renaissance Musicians.*

b. Accomplishments

- Project web site at humanhistoryproject.ca
- Conducted three major experiments
- Four conference presentations: RSA (2014), AMS/SMT (2014), RSA (2015), and IAML/IMS (2015)
- Two workshops held at the Centre for Interdisciplinary Research on Music Media and Technology (CIRMMT) in Montreal on 21 June 2013 and 12 December 2015.
- Seven invited talks (three in Montreal, one in the USA, and two in Japan).
- Training of several graduate students

c. Audiences

The audiences for this project are musicologists, music librarians, Renaissance scholars, digital humanists, and historians in general. Students in courses across the humanities will benefit from the ability to ask queries across disciplines.

d. Evaluations

Evaluations of our system have been accomplished throughout the project by software testing, software quality evaluations, and several formal experiments using the system.

e. Continuation of the Project

The continuation of the project will be realized through applications to several research grants, publicizing the project, gathering other scholars to the project, and encouraging our graduate students to work on the project.

In November 2015, with Fujinaga as the principal investigator, we have applied to the Quebec government (FRQSC) for a 4-year research grant ($640,000CAN) entitled: "Music Information, Research, and Infrastructure (MIRAI)". The study of prosopography of musicians is one of the five major themes of this infrastructure grant.

We have submitted a paper for the Sixteenth Century Society and Conference 2016: 'Cultural networks in the Renaissance: methodological challenges' to be held in Bruges, Belgium, 18–20 August 2016. We also plan to submit another paper for the 2nd International Workshop on Computational History and Data-Driven Humanities, to be held in Dublin, 25 May 2016.

f. Long Term Impact

The long-term impact of this project will be realized through continual development of the system and participation of other scholars to test and use the system. As more data is processed by the system, its utility should prove invaluable, as people will be able to search the database for historical events and discover network of people's relationships in the past.

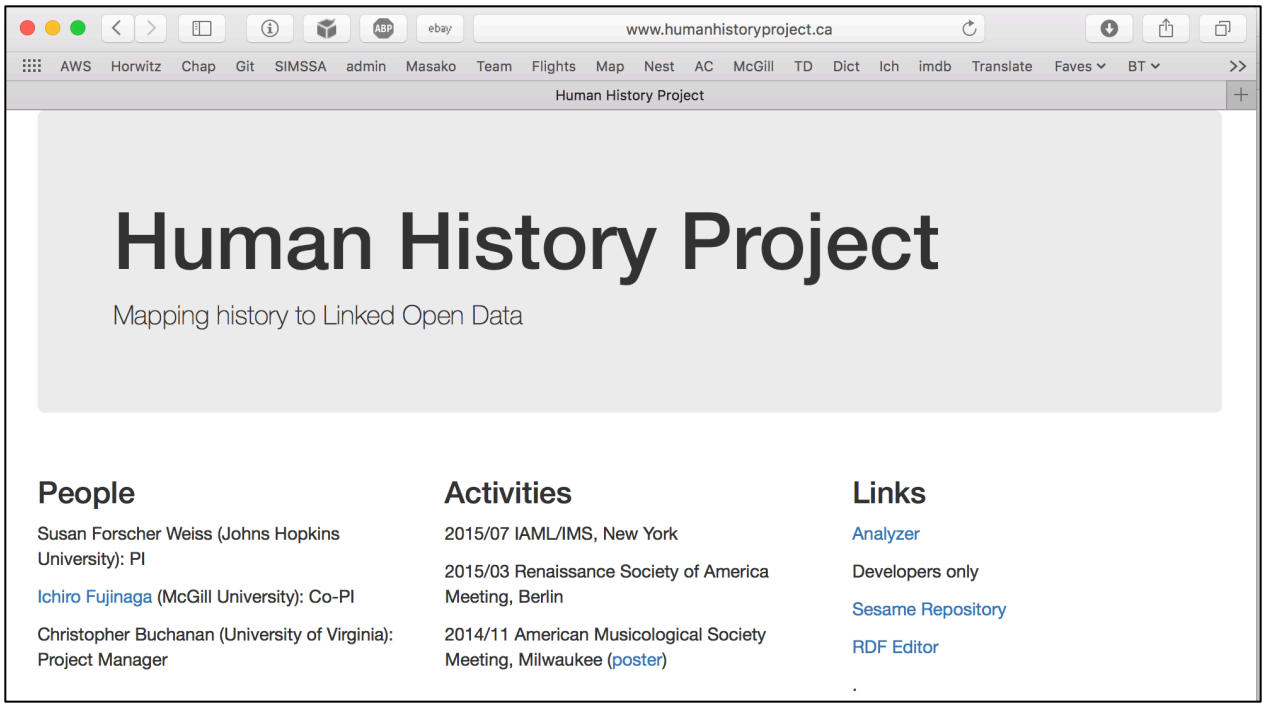
g. Grant Products

- Project web site at humanhistoryproject.ca
- Conference presentations: abstracts, a poster, and several Power Point presentations

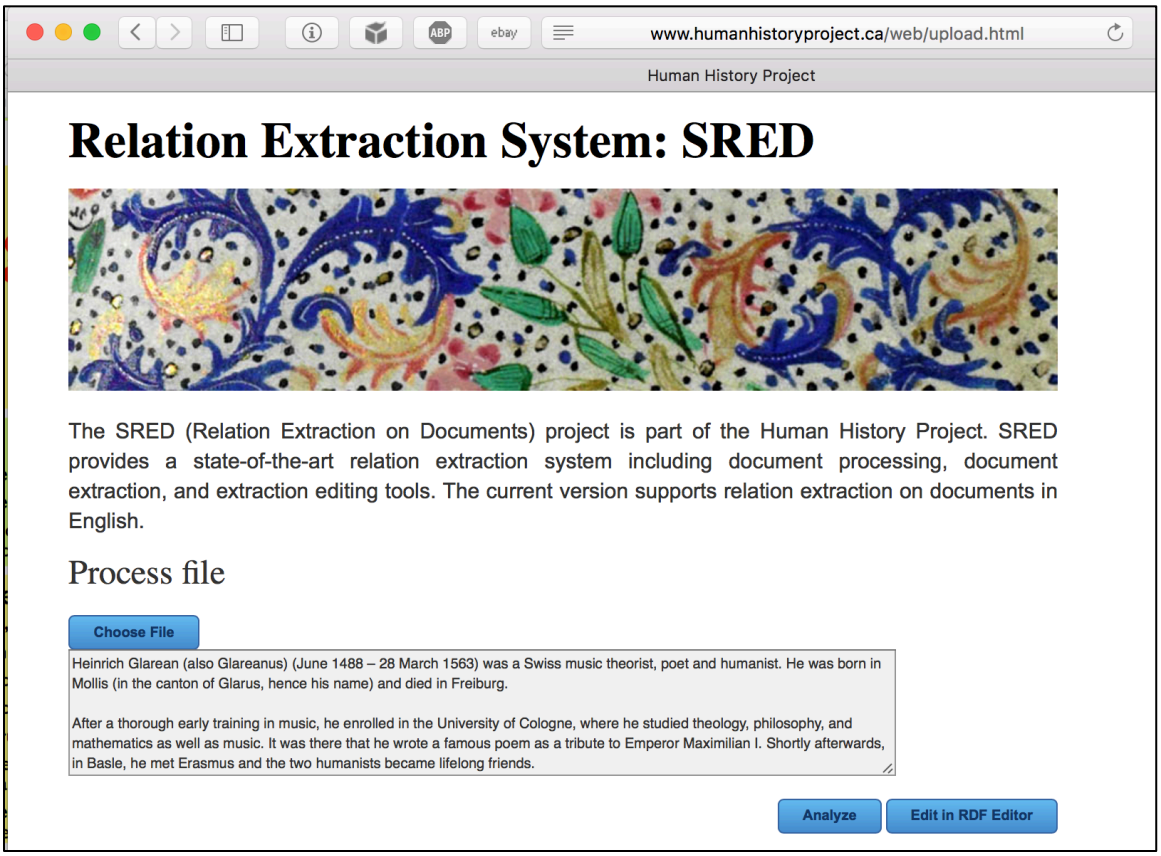
3. Appendix

- A screenshot of the website: The Home page
- A screenshot of the website: The Analyzer page
- A screenshot of the website: The RDF Editor page
- Our abstract in CIM Conference Proceedings: 1/2
- Our abstract in CIM Conference Proceedings: 2/2
- Our poster at AMS/SMT 2014
- CIRMMT Workshop #1 The Science and Technology of Music website (2013)
- CIRMMT Workshop #2 The Science and Technology of Music website (2013)
- Slides from our Presentation 1/4
- Slides from our Presentation 2/4
- Slides from our Presentation 3/4
- Slides from our Presentation 4/4

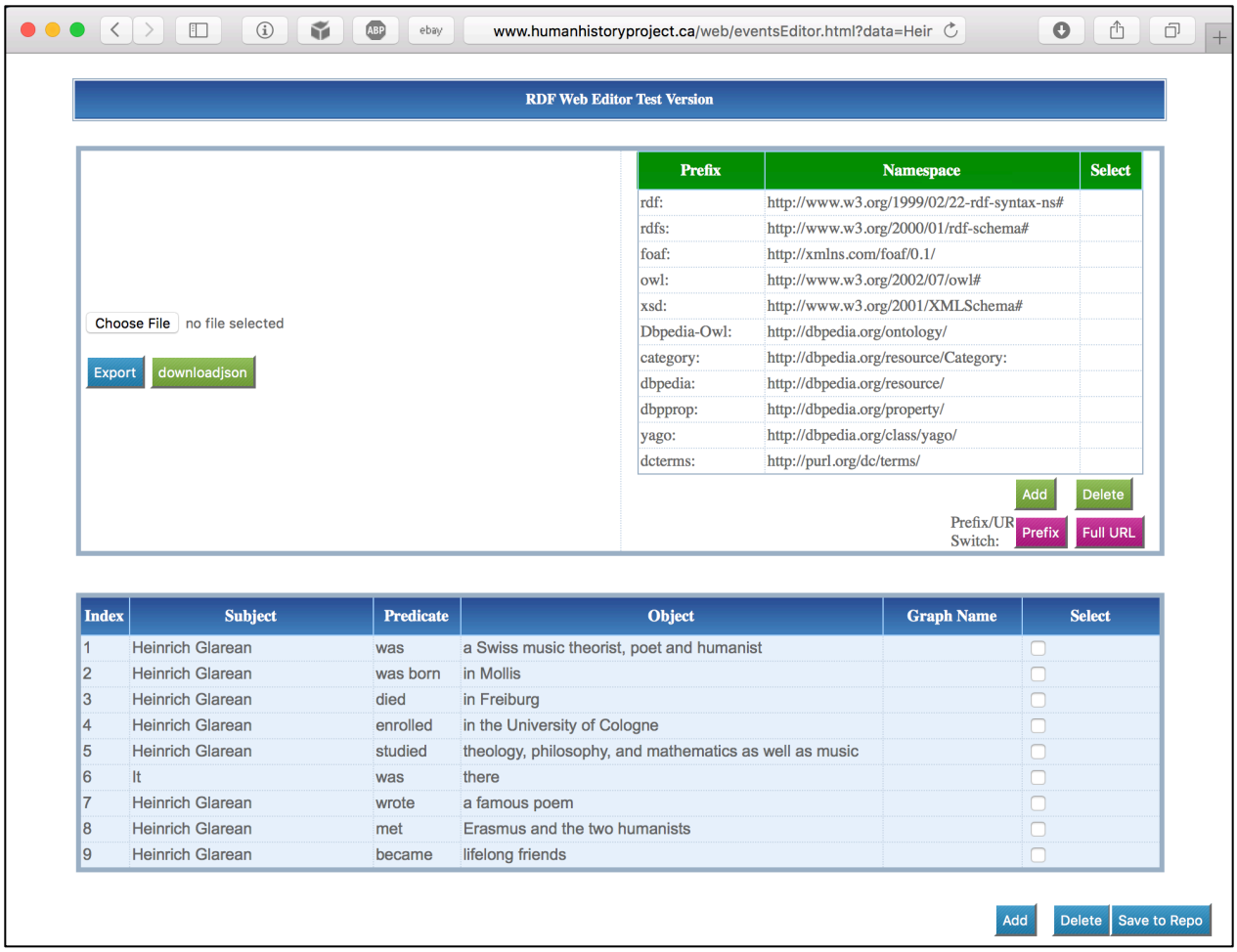
Appendix



A screenshot of the website: The Home page



A screenshot of the website: The Analyzer page



A screenshot of the website: The RDF Editor page

Imagining the musical past: Creating a digital prosopography of Renaissance musicians

Susan Forscher Weiss (Peabody Conservatory of Music, The Johns Hopkins University)
Ichiro Fujinaga (McGill University)

Background in Music History, German & Romance Languages and Literature
Background in Music Technology

Aims

We imagine the music of the past by attempting to perform in as authentic a manner as possible: by building close models of the musical instruments, by studying drawings and paintings, and by close reading of the contemporary treatises. We can also enhance our image of the past musical world by studying the people and their social or professional network based on multiple sources of information. Such a study is called prosopography. By gathering lots of small bits of information, historians can imagine and make sense of the past (Collingwood 1946; Wiseman 1994).

Main Contribution

Currently, building a database of people's relationships is costly and time consuming because it can only be done manually. This project aims to create, automatically, an economically feasible digital prosopographical database of Renaissance musicians by exploiting the ever-increasing availability of historical documents, recent improvements in optical character recognition (OCR) and natural language processing (NLP) technologies, and the emerging data structure called Linked Open Data.

We are creating a framework that can answer questions not easily answered by Google-like searches or traditional means. For example, which printers in Venice in the 1530s were publishing books of music? Which foreign musicians visited Venice in 1538? Did composer A and composer B live in Venice in 1538? Were there musicians working in Venice from 1535–1540 who performed music by both of these composers? Who were the musical instrument makers there in those years?

The possibility of digitizing prosopography has been discussed since the late 1980's (e.g., Bulst 1989). One of the earliest examples of digital prosopography is the *Prosopography of the Byzantine Empire I* (Bradley & Short 2001), produced on CD-ROM. The largest example of digital prosopography is the *Oxford Dictionary of National Biography*. It is a monumental work, which took over 10 years (1992–2004) to complete at a cost of over £25million (Harrison 2004). More recent digital prosopography projects include the Berkeley Prosopography Services (http://berkeleyprosopography.org), the Prosopographie des Chantres de la Renaissance (http://ricercar.cesr.univ-tours.fr/3-programmes/PCR/), and the London Lives Project The latter project, which contains over 3 million names, scanned microfilm sources and entered texts manually (double keyed). The names, occupations, places, and dates were

marked up using "a combination of automated and manual processes."[1] In our project we are further automating the process by using state-of-the-art OCR and NLP technologies. This should drastically reduce the cost compared to previous projects. Because of errors as a result of imperfections in OCR and NLP technologies we plan to deploy crowd- or expert-sourcing techniques for corrections.

The main types of information we are interested in extracting are named entities (person, place, organization, etc.), events, and relationships between named entities and events. GATE (General Architecture for Text Engineering) is our main NLP tool but we are also working with Stanford NLP software (http://nlp.stanford.edu/software/), UIMA (Unstructured Information Management Applications; http://uima.apache.org), and REEL (RElationship Extraction Learning framework; http://reel.cs.columbia.edu) in an effort to improve the results. Nevertheless, because the state-of-the-art NLP technology is imperfect for named entity and events extractions, as well as for the relationship extractions, we have developed a JavaScript-based online editor to correct errors. The results are stored using the quad RDF (Resource Description Framework) format, which then can be searched via SPARQL, a query language for RDF.

We have experimented with the named-entity extraction of the GATE system using biographical entries on ten Renaissance composers from three different sources: Wikipedia, Oxford Music Online, and the 1911 edition of Grove's Dictionary of Music and Musicians. The total of 5,441 entities were extracted with the accuracy of 99.24% precision and 98.9% recall. It should be noted, however, that it took over three hours to manually verify and correct the output from the thirty articles; confirming the need for efficient and economical means of correction.

Implications for Musicological Interdisciplinarity

Even though this project concentrates on musicians of the Renaissance, its model can be applied to other time periods and disciplines. Musicians' lives intersect with artists, writers, clerics, patrons, printers, etc. Combining networks will aid in determining circles of influence and patterns of patronage essential to imagining the past for scholars, teachers, and students of Renaissance culture. As more historical documents are digitized and as the automatic natural language processing improves, a wealth of information that was available but extremely difficult to extract can be more easily retrieved. In fact, this is a pilot project for a more ambitious Human History Project, which aims to create a database for all documented human history, finding references to every individual in the past and able to query any information about them and their relationships with others or to put it another way: "creating a Facebook of the past."

[1] http://www.londonlives.org/static/Project.jsp

Digital Prosopography of Renaissance Musicians

Discovery of Social and Professional Network

Susan Forscher Weiss
Musicology and German & Romance Languages and Literature
The Johns Hopkins University
Robert Lehman Visiting Professor
Harvard Center for Italian Renaissance Studies, Villa i Tatti
sweiss@jhu.edu

Ichiro Fujinaga
Music Technology Area, Department of Music Research
Schulich School of Music, McGill University
Centre for Interdisciplinary Research in Music Media and Technology
ich@music.mcgill.ca

Goals

- To create a social and professional network of Renaissance musicians
- To create a database to be able to study the network
- To make connections and discover relationships or answer questions related to dates, geographies, professions, etc.

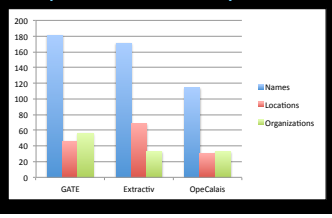
Example queries: Not easily answered by Google, Wikipedia, and other traditional methods

- Which music printers were in business in 1481 in Florence?
- Which composers were residing in Florence in 1481?
- Which composers visited Florence in 1481?
- Which trumpeters were active in Florence between 1481–86?
- What events took place in 1481 in Florence that required musical performance?
- What pieces requiring trumpets were performed in Florence in December 1481?

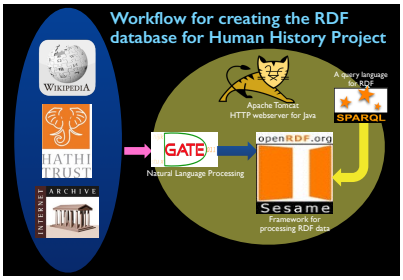
Solutions: Three major tools

- Natural Language Processing (NLP)
  - Named-entity extraction
  - Events extraction
- Linked Open Data
  - RDF (Resource Description Framework): a data model
  - SPARQL: a query language for RDF
- Crowdsourcing

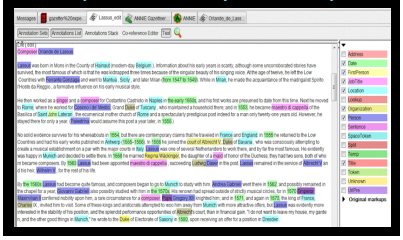
Comparison of Named-Entity Extractors



Workflow for creating the RDF database for Human History Project



An example of named-entity extraction by GATE



THE PEABODY INSTITUTE OF THE JOHNS HOPKINS UNIVERSITY

NATIONAL ENDOWMENT FOR THE HUMANITIES
OFFICE OF DIGITAL HUMANITIES

Named-entity extraction: Experiment I

- Four Renaissance composers' entries in Wikipedia
- Using the default Gazetteer (a dictionary)
  - 90.25% precision (the extracted entities were correctly identified)
  - 65.33% recall (the entities in the document that were found)
- Editing the Gazetteer (~15 minutes / article)
  - 98.39% precision; 91.86% recall
- Fixing the problem with plurals with the Morphological Analyzer (finding root form of words)
  - 98.45% precision; 98.45% recall

Named-entity extraction: Experiment II

- Ten composers x three sources (30) minus the four Wikipedia articles from Experiment 1: 26 articles
- Using the modified Gazetteer and Morphological Analyzer from Experiment 1
  - 99.24% precision; 98.9% recall
- Correction time of the 5,441 entities extracted
  - Average of 3 sec./entity to correct
  - 240 min. to correct 26 articles or about 10 min. / article

Next steps

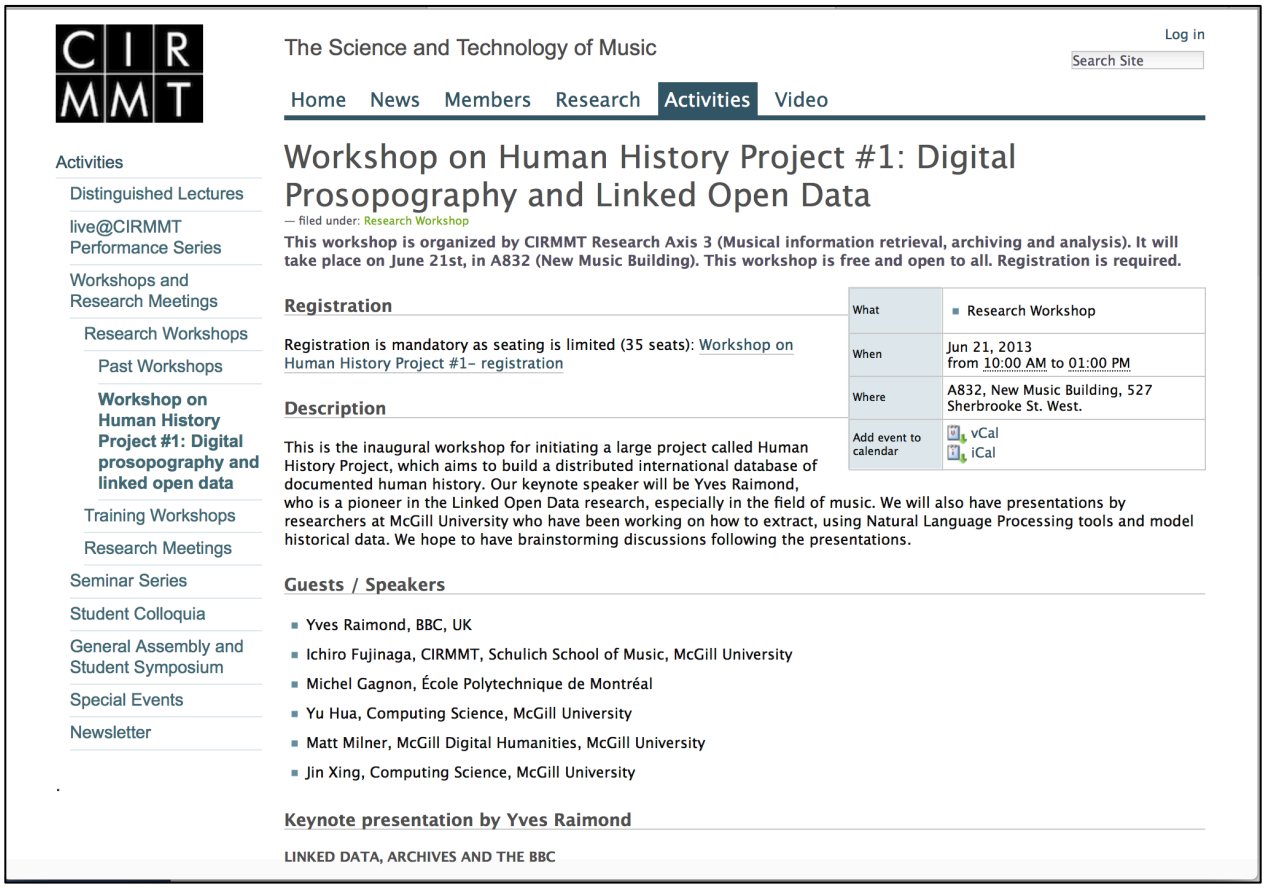
- Extract relationships between named entities (e.g., REEL)
- Create web interface to correct relationships
- Create web interface to query the network of relationships

Acknowledgements

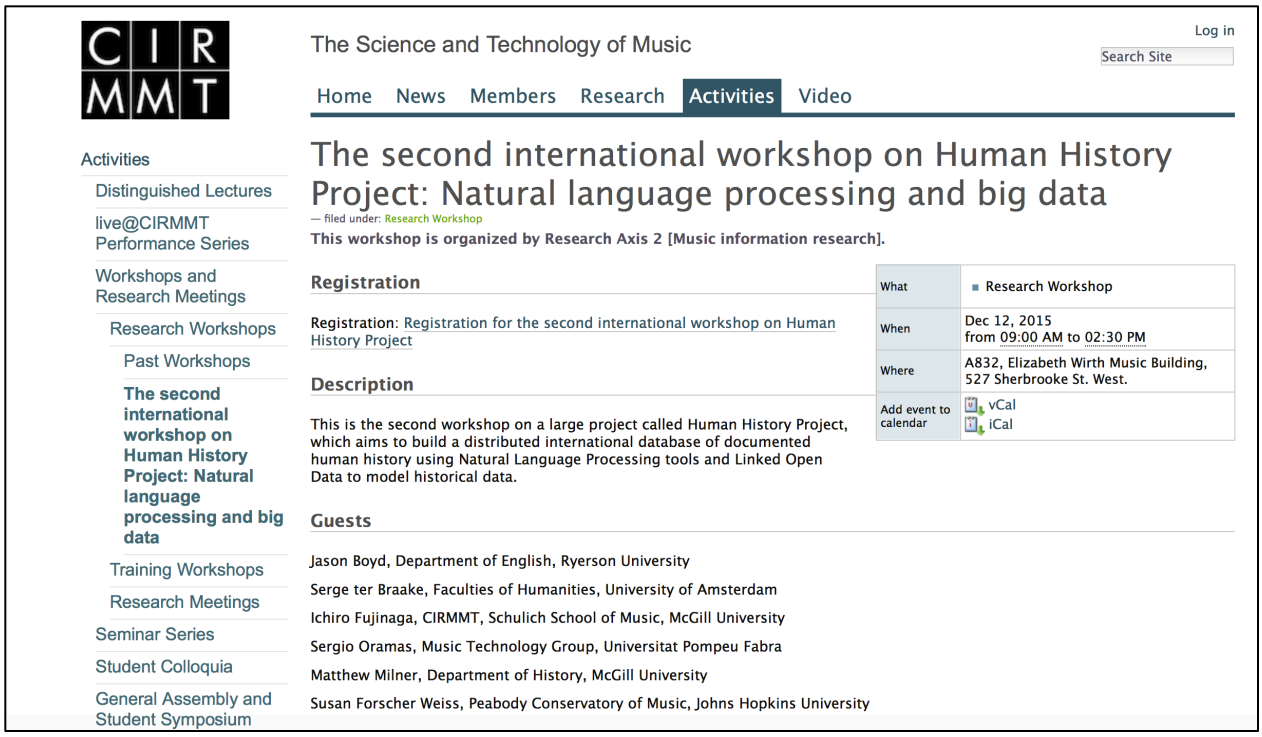
We would like to acknowledge the wonderful work by Christopher Buchanan on this project. Other team members include: Haoyuan Ji, Peng Jianxiang, Ankit Sarwal, and Chengyuan Zhang. This research has been supported by the NEH Digital Humanities Start-up Grant and the Social Sciences and Humanities Research Council of Canada.

CIRMMT Centre for Interdisciplinary Research in Music Media and Technology · McGill · DDMAL DISTRIBUTED DIGITAL MUSIC ARCHIVES & LIBRARIES LAB · Social Sciences and Humanities Research Council of Canada · Conseil de recherches en sciences humaines du Canada · Canada

Our poster at the AMS/SMT 2014



Workshop #1 website (2013)



Workshop #2 website (2015)

Slides from our presentations (1/4)

Slides from our presentations (2/4)

Slides from our presentations (3/4)

LODE
Linking Open Description of Events

"… an ontology for publishing descriptions of historical events as Linked Data, and for mapping between other event-related vocabularies and ontologies."


Summary of Terms

This vocabulary defines one class and 7 properties .

Term Name	Type	Definition
Event	class	"Something that happened," as might be reported in a news article or explained by a historian.
atPlace	property	a named or relatively specified place that is where an event happened.
atTime	property	an abstract instant or interval of time that is when an event happened.
circa	property	an interval of time that can be precisely described using calendar dates and clock times.
illustrate	property	an event illustrated by some thing (typically a media object)
inSpace	property	an abstract region of space (e.g. a geospatial point or region) that is where an event happened.
involved	property	a (physical, social, or mental) object involved in an event.
involvedAgent	property	an agent involved in an event.

Representing events in RDF

- Statement 123: Anna Maria Mozart gives birth to a son
- Statement 124: Statement 123 occurred in 1756
- Statement 125: Statement 124 was written in Wikipedia
- Statement 126: Statement 124 was written on 2010/11/19
- Statement 127: Statement 123 was written in a book
- Statement 128: Statement 123 took place in Salzburg

Provenance Ontologies

- Open Provenance Model
- Provenir Ontology
- Provenance Vocabulary
- etc.
- Provenance Vocabulary Mapping (W3C)

Other tools to try

- Apache OpenNLP
- Stanford NLP Tools
- AFNER: Named Entity Recognition (Australia)
- FRED: "a tool for automatically producing RDF/OWL ontologies and linked data from natural language sentences" (Italy)
- Apache Stanbol: "… provides a set of reusable components for semantic content management"

Existing "structured" data sources

- Freebase / Knowledge Graph
- DBpedia
- Yago (Max Planck Institute)
- *Need to add provenances*
- VIAF (Virtual International Authority Files): unique ID/URI
  - J. S. Bach URI: (http://viaf.org/viaf/12304462/)
- WorldCat Identities (30 million names)
  - 100 billion total (10^{11})
- Getty Thesaurus of Geographic Names (Montréal)

Event extraction applications

- Personalized news systems
- Risk analysis
- Decision-making support tools
- Biological event detectors from academic papers
- Algorithmic stock trading

Event extraction research

- "On-line news event and tracking" (1998)
- "A system for news event detection" (2003)
- "Real-time news event extraction for global crises monitoring" (2008)
- "An overview of event extraction from text" (2011). *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web* (DeRiVE 2011)
- "A survey of techniques for event detection in Twitter" (2013)

Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web

- Started in 2011
- "The goal of this workshop is to bring together those different areas in the recent surge of research on the use of events as a key concept for representing and organizing knowledge on the Web." (http://derive2013.wordpress.com)

HHP: Multidisciplinary Research

- History
- Linguistics
- Computer Science
- Information Science
- Geography
- Philosophy ("*what sort of things are events?*")
- Onomastics: Study of proper names
- Toponymy: Study of place names
- Hydronymy: Study of a proper name of a body of water

HHP: Future Plans

- Suggestion for workshops
- Build an infrastructure at McGill?
- Grant applications
- Iterative improvements over the years (next 50–100 years)
- Logo design / designers?

Basic Entity Categories

- Entity names (ENAMEX): Organization, person, location, place names (lake, river, mountain)
- Temporal expression (TIMEX): Date, time
- Number expressions (NUMEX): Money, percent, age, weight, distance

Named-entity Recognition

- The term "named-entity recognition (NER) first used in a paper by Ralph Grishman (Computer Science, NYU) Proceedings of the workshop on Human Language Technology (held at Plainsboro, NJ, March 8-11, 1994)
- NER was developed as part of the evaluation for Message Understanding Conference, (MUC) which began in 1987. As part of the evaluation suite MUC-6 using TREC corpora
- MUC (began in 1987) have been organized by NRAD, the Research Development, Testing, and Evaluation division the Naval Command, Control and Ocean Surveillance Center with the support of DARPA (the Defense Advanced Research Projects Agency)
- Message Understanding Conferences were forum "to assess and to foster research on the automated analysis of military messages…" (Grishman and Sundheim 1996, 466)

History of Message Understanding Conferences (MUC)

- "Although called 'conferences', the distinguishing characteristic of the MUCs are not the conferences themselves, but the evaluations to which participants must submit in order to be permitted to attend the conference." (Grishman and Soundheim 1996, 466)
- "MUC-1 (1987) was basically exploratory…. By MUC-2 (1989), the task" was to "fill a template with information about the event, such as the type of event, the agent, the time and place, the effect, etc."

Slides from our presentations (4/4)